Published online 21 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D983–D986***
doi:10.1093/nar/gks1099

LncRNADisease: a database for long-non-coding RNA-associated diseases

Geng Chen[1,2,3], Ziyun Wang[2], Dongqing Wang[2], Chengxiang Qiu[2], Mingxi Liu[4], Xing Chen[4], Qipeng Zhang[2], Guiying Yan[4,*] and Qinghua Cui[1,2,3,*]

[1]Department of Biomedical Informatics, School of Basic Medical Sciences, [2]MOE Key Lab of Cardiovascular Sciences, [3]Institute of Systems Biomedicine, Peking University, 38 Xueyuan Road and [4]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China

Received July 5, 2012; Revised October 11, 2012; Accepted October 21, 2012

ABSTRACT

In this article, we describe a long-non-coding RNA (lncRNA) and disease association database (LncRNADisease), which is publicly accessible at http://cmbi.bjmu.edu.cn/lncrnadisease. In recent years, a large number of lncRNAs have been identified and increasing evidence shows that lncRNAs play critical roles in various biological processes. Therefore, the dysfunctions of lncRNAs are associated with a wide range of diseases. It thus becomes important to understand lncRNAs' roles in diseases and to identify candidate lncRNAs for disease diagnosis, treatment and prognosis. For this purpose, a high-quality lncRNA–disease association database would be extremely beneficial. Here, we describe the LncRNADisease database that collected and curated approximately 480 entries of experimentally supported lncRNA–disease associations, including 166 diseases. LncRNADisease also curated 478 entries of lncRNA interacting partners at various molecular levels, including protein, RNA, miRNA and DNA. Moreover, we annotated lncRNA–disease associations with genomic information, sequences, references and species. We normalized the disease name and the type of lncRNA dysfunction and provided a detailed description for each entry. Finally, we developed a bioinformatic method to predict novel lncRNA–disease associations and integrated the method and the predicted associated diseases of 1564 human lncRNAs into the database.

INTRODUCTION

A surprising finding in human transcriptome analysis is that protein-coding sequences only account for a small portion of the genome transcripts (1). The majority of the human genome transcripts are non-coding RNAs, in particular, long-non-coding RNAs (lncRNAs) (2). Normally, lncRNAs tend to be less conserved across species and often show low expression levels and high tissue specificity (3–5). Thus, at the time they were first found, lncRNAs were often considered to be transcriptional noise (5). In recent years, accumulating studies have revealed that a number of lncRNAs are not transcriptional noise but have important functions, for example, affecting gene transcription, targeting RNA polymerase II, regulating splicing and taking part in epigenetics (6). Moreover, according to the theory of competing endogenous RNA (7), lncRNAs may functionally interact with a broad range of RNA molecules through competitively binding with microRNA (miRNAs), suggesting that lncRNAs may have critical roles in a wide range of biological processes. Previous studies produced a large amount of lncRNA-related data, including sequences, expression profiles and functions. Therefore, arranging and annotating these data are important to better understand lncRNAs. Several databases for lncRNAs indeed provide helps in studying lncRNAs (8–10). For example, NRED is a database for lncRNA expression data (10). The lncRNAdb database provides detailed lncRNA information, including sequences, functions, expressions, associated proteins and cellular locations (8). Although the NONCODE database is not specific to lncRNA, it curates the sequences, functions, expressions and cellular location of lncRNAs in the third version (NONCODE v3.0) (9).

*To whom correspondence should be addressed. Tel: +86 10 82801585; Fax: +86 10 82801001; Email: cuiqinghua@hsc.pku.edu.cn
Correspondence may also be addressed to Guiying Yan. Tel: +86 10 62574529; Fax: +86 10 62561939; Email: yangy@amss.ac.cn

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

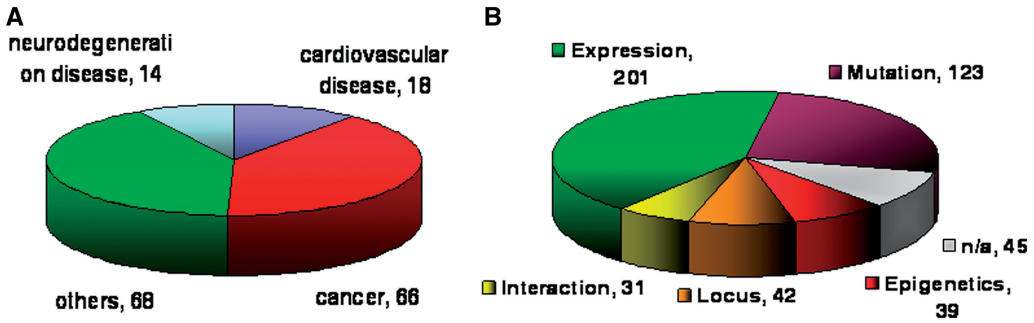

Figure 1. Statistics and distributions of diseases (**A**) and dysfunction types (**B**) of lncRNAs in the LncRNADisease database.

More recently, researchers have attempted to understand the relationships between lncRNAs and diseases. Studies have reported that lncRNA dysfunctions are associated with a broad range of diseases (5), including cancers (11), cardiovascular diseases (12) and neurodegeneration diseases (13). For example, lncRNA *PCA3* is a highly prostate cancer-specific molecules and a *PCA3* score has the potential to be a biomarker for prostate cancer aggressiveness (14). The up-regulation of the lncRNA *HOTAIR* is an independent prognostic factor of tumor recurrence in hepatocellular carcinoma patients after liver transplantation (15). A study confirmed the high specificity and sensitivity of lncRNA *UCA1* from urinary sediments in the diagnosis of bladder cancer, suggesting that *UCA1* is a potential biomarker for bladder cancer diagnosis (16). Godinho *et al.* (17) revealed that the lncRNA *BCAR4* can be a potential target for antiestrogen-resistant breast cancer treatment because its forced expression in breast cancer cells leads to cell proliferation in the presence of various antiestrogens and in the absence of estrogen. The above studies indicate that lncRNAs may help to understand diseases and help to find potential molecules in disease diagnosis, treatment and prognosis. Therefore, the study of lncRNA–disease associations is becoming one of the most important topics of lncRNAs and diseases. For this reason, a high-quality lncRNA–disease association database will be helpful in studying the roles of lncRNAs in diseases but is still not available. To build such a database, we manually curated lncRNA–disease relations experimentally reported in the literature and created a database, LncRNADisease. We included detailed annotation information for each entry. Moreover, we curated and annotated experimentally supported lncRNA interacting partners. In addition, we developed a bioinformatic method to predict novel lncRNA–disease associations and integrated this method and its predicted results into the database.

DATA SOURCES AND IMPLEMENTATION

First, we downloaded PubMed data, information on non-protein-coding RNA genes, and data on gene–PubMed associations from the National Center for Biotechnology Information. Second, we curated the data manually and retrieved lncRNA–disease pairs. All lncRNA–disease pairs were double-checked by different researchers. Hyperlinks to the original articles in PubMed database were provided. We also annotated the sequence and species information. We further normalized the names of lncRNAs and diseases. In total, we curated 166 diseases, of which cancer (39.8%), cardiovascular disease (10.8%) and neurodegeneration disease (8.4%) were the top three classes (Figure 1A). Moreover, we provided detailed descriptions for the associations of lncRNAs and diseases and curated the dysfunction type for each entry. For example, if an entry's dysfunction evidence is derived from expression data, the dysfunction type of this entry will be considered as 'Expression'. The distribution of the dysfunction type is shown in Figure 1B. Aside from lncRNA–disease association data, we also curated experimentally supported lncRNA interactions and cataloged the interactions according to the interacting molecules and the characteristics of the interactions. For example, at the RNA level, lncRNAs may interact with proteins (18), RNAs (19), lncRNAs (20) and miRNAs (21). Their interactions may be binding, regulation and co-expression. At the DNA level, promoters of lncRNA genes may bind with transcription factors (TFs) and be regulated by TFs (22).

All data were organized in the 'LncRNADisease' database using SQLite, a lightweight database management system. The website was developed based on Django, a Python web framework. The database is available at http://cmbi.bjmu.edu.cn/lncrnadisease.

PREDICTING NOVEL LNCRNA–DISEASE ASSOCIATIONS

LncRNADisease was designed not only as a resource for experimentally supported lncRNA–disease association data, but also as a platform for predicting novel lncRNA–disease associations. In this study, we present a method to predict novel lncRNA–disease associations based on the genomic context of a given lncRNA. We previously showed that miRNAs located closely to each other in the genome (particularly miRNAs within 2 kb) and tend to be associated with similar diseases (23,24). Here, we investigated whether or not lncRNAs tend to be associated with a similar disease as their genomic neighbor genes. Thus, we identified the protein-coding genes and miRNAs within 2 kb nts of any lncRNA with

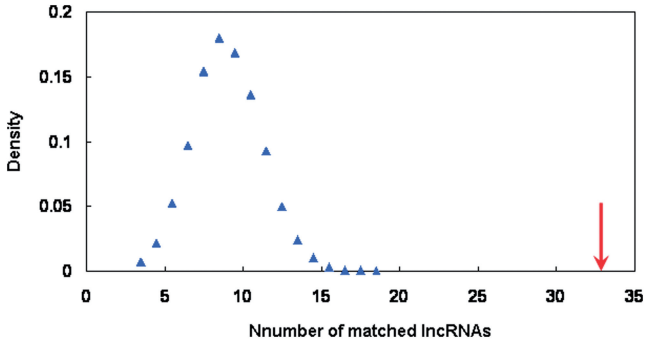

Figure 2. Significance of lncRNAs sharing a common disease with their neighbor genes/miRNAs. Blue triangles indicate the distributions of numbers of lncRNAs associated with the same disease as their neighbor genes/miRNAs in random cases. The red arrow indicates the real number of lncRNAs associated with the same disease as their neighbor genes/miRNAs.

reported disease associations. We then identified the lncRNAs with the same associated disease as the neighbor genes/miRNAs. We found 33 lncRNAs associated with the same disease as their neighbor genes/miRNAs. To evaluate the significance, we randomly re-permuted the disease associated with the lncRNAs for 10 000 times and counted the number of lncRNAs associated with the same associated disease as their neighbor genes/miRNAs. As a result, none of the counts was greater than 33 and the expected number was 9, indicating that lncRNAs and their neighbor genes/miRNAs tend to be associated with the same disease ($P < 1 \times 10^{-4}$, randomization test; Figure 2). This result suggests that we can predict potential-associated disease for lncRNA through the disease associated with its neighbor genes/miRNAs. Based on the above observation, we developed a tool to predict novel lncRNA–disease associations and identified potential-associated diseases for all lncRNAs identified in the human genome using this tool. Finally, we integrated the tool and the predicted results into the LncRNADisease database.

QUERYING THE DATABASE

We provide users several ways to query the LncRNADisease database. First, users can browse the LncRNADisease by lncRNA names or disease names. When clicking one lncRNA or disease in the 'Browse' page, LncRNADisease will return a list of matched entries. Second, we provide a 'fuzzy search' function for the entries by the full or partial names of lncRNAs or diseases in the 'Search' page. The 'Search' is case insensitive. We also provide a page for tools to predict novel lncRNA–disease associations. Moreover, all data in the database, including lncRNA–disease associations, predicted lncRNA–disease associations and lncRNA interactions, can be downloaded. The users can also submit novel data into the database. In addition, a detailed tutorial for the usage of the database is available in the 'Help' page.

FUTURE EXTENSIONS

The LncRNADisease database represents the first step in this project. Further extensions will be developed. The LncRNADisease database will update the experimentally supported lncRNA–disease association data every 2 months. Meanwhile, some new tools for analyzing lncRNA–disease association data is being developed and will be integrated into the LncRNADisease database in the future. For example, we are developing expression profile- and interacting partner-based methods to predict novel lncRNA–disease associations and expect to integrate these methods into the database in the near future.

DISCUSSION AND CONCLUSION

Increasing studies have shown that lncRNAs have important functions and are associated with a broad range of diseases. LncRNAs are becoming novel potential molecules for disease diagnosis, treatment and prognosis. In this article, we describe an lncRNA and disease association database, LncRNADisease. The LncRNADisease database integrated several types of data, such as experimentally supported lncRNA–disease association data, experimentally supported lncRNA interaction data and predicted lncRNA–disease association data. Moreover, we developed a bioinformatic method to predict potential-associated disease for a novel lncRNA based on its genomic context and integrated this method into LncRNADisease.

The important roles of lncRNAs in disease are attracting more biomedical researchers. Therefore, more experimentally supported lncRNA–disease associations are expected to be published in the future and these data will be integrated into the LncRNADisease database. More importantly, although thousands of lncRNA have been identified, only a limited number of lncRNAs have been reported to be associated with diseases. It is increasingly needed to predict potential-associated diseases for lncRNAs through bioinformatic methods. Therefore, another major aim of LncRNADisease is to develop and integrate more bioinformatic methods for analyzing and predicting lncRNA–disease associations. Finally, we believe that LncRNADisease is useful for the studies of lncRNAs and diseases, and will provide more helps in this topic when it integrates more data and tools in the future.

FUNDING

National Basic Research program of China [2012CB517500]; National Natural Science Foundation of China [31000585 and 11021161]. Funding for open access charge: National Basic Research program of China [2012CB517500].

Conflict of interest statement. None declared.

REFERENCES

1. Bertone,P., Stolc,V., Royce,T.E., Rozowsky,J.S., Urban,A.E., Zhu,X., Rinn,J.L., Tongprasit,W., Samanta,M., Weissman,S.

et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.
2. Kapranov,P., Cheng,J., Dike,S., Nix,D.A., Duttagupta,R., Willingham,A.T., Stadler,P.F., Hertel,J., Hackermuller,J., Hofacker,I.L. *et al*. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.
3. Mercer,T.R., Dinger,M.E., Sunkin,S.M., Mehler,M.F. and Mattick,J.S. (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl Acad. Sci. USA*, **105**, 716–721.
4. Pauli,A., Valen,E., Lin,M.F., Garber,M., Vastenhouw,N.L., Levin,J.Z., Fan,L., Sandelin,A., Rinn,J.L., Regev,A. *et al*. (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res*., **22**, 577–591.
5. Ponting,C.P., Oliver,P.L. and Reik,W. (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.
6. Managadze,D., Rogozin,I.B., Chernikova,D., Shabalina,S.A. and Koonin,E.V. (2011) Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol. Evol*., **3**, 1390–1404.
7. Salmena,L., Poliseno,L., Tay,Y., Kats,L. and Pandolfi,P.P. (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, **146**, 353–358.
8. Amaral,P.P., Clark,M.B., Gascoigne,D.K., Dinger,M.E. and Mattick,J.S. (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res*., **39**, D146–D151.
9. Bu,D., Yu,K., Sun,S., Xie,C., Skogerbo,G., Miao,R., Xiao,H., Liao,Q., Luo,H., Zhao,G. *et al*. (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res*., **40**, D210–D215.
10. Dinger,M.E., Pang,K.C., Mercer,T.R., Crowe,M.L., Grimmond,S.M. and Mattick,J.S. (2009) NRED: a database of long noncoding RNA expression. *Nucleic Acids Res*., **37**, D122–D126.
11. Spizzo,R., Almeida,M.I., Colombatti,A. and Calin,G.A. (2012) Long non-coding RNAs and cancer: a new frontier of translational research? *Oncogene*, **31**, 4577–4587.
12. Congrains,A., Kamide,K., Oguro,R., Yasuda,O., Miyata,K., Yamamoto,E., Kawai,T., Kusunoki,H., Yamamoto,H., Takeya,Y. *et al*. (2012) Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B. *Atherosclerosis*, **220**, 449–455.
13. Johnson,R. (2012) Long non-coding RNAs in Huntington's disease neurodegeneration. *Neurobiol. Dis*., **46**, 245–254.
14. van Poppel,H., Haese,A., Graefen,M., de la Taille,A., Irani,J., de Reijke,T., Remzi,M. and Marberger,M. (2012) The relationship between Prostate CAncer gene 3 (PCA3) and prostate cancer significance. *BJU Int*., **109**, 360–366.
15. Yang,Z., Zhou,L., Wu,L.M., Lai,M.C., Xie,H.Y., Zhang,F. and Zheng,S.S. (2012) Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation. *Ann. Surg. Oncol*., **18**, 1243–1250.
16. Zhang,Z., Hao,H., Zhang,C.J., Yang,X.Y., He,Q. and Lin,J. (2012) [Evaluation of novel gene UCA1 as a tumor biomarker for the detection of bladder cancer]. *Zhonghua Yi Xue Za Zhi*, **92**, 384–387.
17. Godinho,M., Meijer,D., Setyono-Han,B., Dorssers,L.C. and van Agthoven,T. (2011) Characterization of BCAR4, a novel oncogene causing endocrine resistance in human breast cancer cells. *J. Cell Physiol*., **226**, 1741–1749.
18. Pasmant,E., Sabbagh,A., Vidaud,M. and Bieche,I. (2011) ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J*., **25**, 444–448.
19. Faghihi,M.A., Modarresi,F., Khalil,A.M., Wood,D.E., Sahagan,B.G., Morgan,T.E., Finch,C.E., St Laurent,G., 3rd, Kenny,P.J. *et al*. (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med*., **14**, 723–730.
20. Clark,M.B. and Mattick,J.S. (2011) Long noncoding RNAs in cell biology. *Semin. Cell Dev. Biol*., **22**, 366–376.
21. Wilusz,J.E., Sunwoo,H. and Spector,D.L. (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev*., **23**, 1494–1504.
22. Koshimizu,T.A., Fujiwara,Y., Sakai,N., Shibata,K. and Tsuchiya,H. (2010) Oxytocin stimulates expression of a noncoding RNA tumor marker in a human neuroblastoma cell line. *Life Sci*., **86**, 455–460.
23. Lu,M., Zhang,Q., Deng,M., Miao,J., Guo,Y., Gao,W. and Cui,Q. (2008) An analysis of human microRNA and disease associations. *PLoS One*, **3**, e3420.
24. Wang,D., Wang,J., Lu,M., Song,F. and Cui,Q. (2010) Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics*, **26**, 1644–1650.